新疆天文台 Taurus 高性能计算系统*

张海龙[1,2]，冶鑫晨[1]，聂俊[1,2]，陈龙飞[1]，托乎提努尔[1]，王杰[1]，崔辰州[3]，李长华[3]，朱艳[1,4]，张萌[1,4]

（1. 中国科学院新疆天文台，新疆 乌鲁木齐，830011；2. 中国科学院射电天文重点实验室，江苏 南京，210008；3. 中国科学院国家天文台，北京，100012；4.中国科学院大学，北京，100049）

摘要：新疆天文台 Taurus 高性能计算系统由 1 个登陆节点、16 个计算节点、2 个 I/O 节点、100TB 高速存储组成。其 CPU 理论双精浮点计算能力 6.7584Tflops，通过 Linpack 测试实际峰值为 6.289Tflops,可提供计算能力为理论值的 93.06%；GPU 理论计算能力 18.72 Tflops，实际测试计算峰值为 14.882Tflops,计算效率为理论能力的 79.5%。系统计算与存储节点均采用 56Gb Infiniband 交换互连，通过 IOZone 测试存储系统，单节点写可达 460MB/s,多节点写可达 800MB/s。系统已在多相滤波及消干扰 GPU 算法加速、蒙特卡罗模拟等领域得到了应用。

关键字：HPC；Lustre；IOZone

中图分类号：P111.5； TP391 文献标识码：A 文章编号：1672-7673（2018）

---

* 基金项目：国家自然科学基金(U1531125)；国家重点基础研究发展计划（973）项目（2015CB857100）；中国科学院青年创新促进会；中国科学院天文台站设备更新及重大仪器设备运行专项经费资助.

收稿日期:2017-09-21;修订日期:2017-10-18

作者简介：张海龙，男，博士. 研究方向：数据密集型研究.Email：zhanghailong@xao.ac.cn

天文学是一门基于观测和模拟的数据密集型科学，暗能量、暗物质、黑洞、宇宙起源、天体起源、生命起源等是天文学研究的前沿重大基础科学问题，解决这些问题的首要方法是观测，其次是利用高性能计算系统对海量真实或仿真数据进行建模和模拟分析，从而进一步揭示宇宙的奥秘。信息技术、高性能计算技术等的高速发展使得用计算机实现理论和实验研究成为可能，在这样的背景下，天文学家通过高性能计算系统进行科学研究成为必然趋势杨哲睿, 高娜, 刘梁. 大规模天文数据分析及多维信息可视化平台的建设和管理[J]. 科研信息化技术与应用, 2015, 6(5):73-83.。中国科学院新疆天文台几十年来致力于天文观测和理论研究，科研内容包括：脉冲星、恒星形成、活动星系核、射电及光学望远镜技术等，为提高数据处理和仿真分析研究效率，结合自身科学研究需求搭建了高性能计算系统，命名为Taurus[1]。

近些年基于图形处理器(Graphic Processing Unit,GPU)的并行计算技术已经成为高性能计算领域的研究热点，利用GPU可以大大加速科学分析、仿真等方面应用程序的运行速度Fan Z, Qiu F, Kaufman A, et al. GPU cluster for high performance computing[C]// Proceedings of the ACM/IEEE SC2004 Conference. 2004: 47-47.。GPU加速计算技术早在2007年由NVIDIA公司推出Kirk D. NVIDIA CUDA software and GPU parallel computing architecture[C]// Proceedings of the 6th international symposium on Memory management. 2007: 103-104.，将计算密集型的任务提交GPU处理，同时CPU依然处理其余任务，可以有效提升数据处理速度。在天文领域，GPU计算框架非常适合天文图像处理、宇宙学大尺度数值模拟、空间目标轨道模拟等，GPU计算框架已经在天文学研究中得到广泛的应用。

高性能计算系统的计算性能来自于多节点的并行计算，节点之间的数据传输、通讯是系统建设的关键Chervenak A, Foster I, Kesselman C, et al. The data grid: towards an architecture for the distributed management and analysis of large scientific datasets[J]. Journal of Network and Computer Applications, 2000, 23(3): 187-200.。Taurus高性能计算系统的建立使得新疆天文台在高性能计算支持上实现了零的突破，在未来的工作中Taurus高性能计算系统将助力于新疆天文台在天体演化模型研究、射电天文多相滤波器、相干及非相干消色散、数值模拟等多方面的科研工作。

1 http://taurus.xao.ac.cn/

1、Taurus 高性能计算系统

1.1 系统拓扑

Taurus 高性能计算系统采用 CPU+GPU 混合架构，目前整个系统由 1 个登陆节点、1 个管理节点、16 个计算节点、2 个 I/O 节点、100TB 高速存储组成。每个计算节点配备了 2 颗 12 核心 Intel Xeon E5-2692 v2 CPU，主频为 2.20GHz，64GB 内存，一个 Nvidia Tesla K20m GPU。计算与 I/O 节点之间通过 56Gb Infiniband 交换机互联，以实现调整数据或消息传递；千兆以太网及 IPMI 网络用于集群系统管理，Taurus 高性能计算系统拓扑结构如图 1。

图 1 Taurus 高性能计算系统拓扑结构图

Fig.1 Topology of Taurus

1.2 计算性能测试

LinpackDongarra J J, Luszczek P, Petitet A. The LINPACK benchmark: past, present and future[J]. Concurrency and Computation: Practice and Experience, 2003, 15(9): 803-820.是国际上最流行的用于测试高性能计算机系统浮点性能的基准程序，也是世界排名 TOP500 超级计算机的标准测试软件。性能测试由多个 64 位双精浮点运算组成，测试一个计算系统每秒可以进行的乘加计算次数（flops）。Linpack 有 3 种基准测试，分别为 Linpack 100、Linpack 1000 以及 HPLinpack Barrett R F, Chan T H F, D'Azevedo E F, et al. Complex version of high performance computing LINPACK benchmark (HPL)[J]. Concurrency and Computation: Practice and Experience, 2010, 22(5): 573-587.。前两种基准测试不适合测试并行计算机集群，本文采用 HPLinpack 对 Taurus 高性能计算系统进行测试。

1.2.1 CPU 性能测试

Taurus 高性能计算系统共有 16 个计算节点，单个计算节点配置如表 1。E5-2692v2 每时钟周期可进行 8 次运算，Taurus 高性能计算系统 CPU 双精度浮点理论计算能力为 2.2*8*24*16= 6758.4Gflops。

表 1 计算节点配置表

Table 1 Compute node configuration

名称	类别	数量	备注
CPU	E5-2692v2	2	单 CPU 核数 12 个、24 线程，30MB 缓存
GPU	Tesla K20M	1	核心数 2496
内存	8G DDR3	8	单节点 64GB 内存
网络	56Gb Infiniband Gigabit		Infiniband 用于数据传输 千兆以太网用于管理
存储 1	300GB	2	本地存储，SAS 接口，10000 转/分
存储 2	200GB	1	挂载的管理节点 opt,用于软件同步
存储 3	100TB	1	挂载的集中式 Lustre 文件系统

HPLinpack 是针对现代并行计算机提出的测试方法，其核心是利用高斯消元法求解一元 *N* 次幂稠密线性代数方程组，测试和评价高性能计算系统的浮点运算性能。Linpack 的 HPL.dat 文件配置如表 2，16 节点 CPU 测试结果如表 3。

表 2 CPU 测试 HPL.dat 配置表

Table 2 HPL.dat configuration table of CPU test

HPLinpack benchmark input file	
Innovative Computing Laboratory, University of Tennessee	
HPL.out	output file name (if any)
6	device out (6=stdout,7=stderr,file)
1	# of problems sizes (N)
341760	Ns
1	# of NBs
208	NBs
1	PMAP process mapping (0=Row-,1=Column-major)
1	# of process grids (P x Q)
16	Ps
24	Qs
16.0	Threshold
1	# of panel fact

1	PFACTs (0=left, 1=Crout, 2=Right)
1	# of recursive stopping criterium
1	NBMINs (>= 1)
1	# of panels in recursion
2	NDIVs
1	# of recursive panel fact
1	RFACTs (0=left, 1=Crout, 2=Right)
1	# of broadcast
3	BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
1	# of lookahead depth
2	DEPTHs (>=0)
2	SWAP (0=bin-exch,1=long,2=mix)
64	swapping threshold
0	L1 in (0=transposed,1=no-transposed) form
0	U in (0=transposed,1=no-transposed) form
1	Equilibration (0=no,1=yes)
8	memory alignment in double (> 0)

表 3 16 节点 CPU 测试结果

Table 3 Test results of 16 CPU nodes

T/V	N	NB	P	Q	Time	Gflops
WC23C2C1	341760	208	16	24	4231.70	6.289e+03

最终测试结果表明，Taurus 高性能计算系统 CPU 双精度浮点实际计算能力为 6.289Tflops，计算效率为 6.289/6.7584=93.06%。

1.2.2 GPU 性能测试

Taurus 高性能计算系统每个节点配备一块 Nvidia Tesla K20m GPU，Tesla K20m 是 Nvidia 推出的 Kepler 架构 GPU，该 GPU 拥有 2496 个 CUDA 核心，核心频率为 706MHz，存储器带宽为 208GB/s，Taurus 高性能计算系统 16 节点 GPU 双精度浮点数理论计算能力为 16*1.17e+03Gflops=18.72Tflops。

本文使用 HPLinpack 对单个节点 GPU 计算性能进行了测试，Linpack 的 HPL.dat 文件配置如表 4，单个 Tesla K20m GPU 不同 Ns 值测试结果如表 5。

表 4 GPU 测试 HPL.dat 配置表

Table 4 HPL.dat configuration table of CPU test

HPLinpack benchmark input file

Innovative Computing Laboratory, University of Tennessee	
HPL.out	output file name (if any)
6	device out (6=stdout,7=stderr,file)
5	# of problems sizes (N)
82544 82897 87936 88192 88448	Ns
1	# of NBs
1024	NBs
0	PMAP process mapping (0=Row-,1=Column-major)
1	# of process grids (P x Q)
1	Ps
1	Qs
16.0	Threshold
1	# of panel fact
0 1 2	PFACTs (0=left, 1=Crout, 2=Right)
1	# of recursive stopping criterium
2 8	NBMINs (>= 1)
1	# of panels in recursion
2	NDIVs
1	# of recursive panel fact
0 1 2	RFACTs (0=left, 1=Crout, 2=Right)
1	# of broadcast
0 2	BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
1	# of lookahead depth
1 0	DEPTHs (>=0)
1	SWAP (0=bin-exch,1=long,2=mix)
192	swapping threshold
1	L1 in (0=transposed,1=no-transposed) form
1	U in (0=transposed,1=no-transposed) form
1	Equilibration (0=no,1=yes)
8	memory alignment in double (> 0)

表 5 单 GPU 测试结果

Table 5 Test results of GPU

Ns Value	Tesla K20M Calculation Value(Gflops)
82544	1.034e+03
82897	1.057e+03

87936	1.069e+03
88192	1.065e+03
88448	9.063e+02

通过多次测试得到了 Ns 值的经验公式，如计算节点的总内存为 *M*，节点的个数为 *N*，系数为 *R*，则：

Ns=ROUND(SQRT(M*N*1024*1024*R/8)/128,0) *128

当 *R* 值在 0.8 到 0.9 之间时 GPU 集群可以得到最高测试结果。测试中 Taurus 高性能计算系统 16 个 GPU 节点最高计算峰值为 14.882Tflops，实际计算效率为 14.882/18.72=79.5%。

1.3 存储系统性能测试

Taurus 采用了 LustreZhao T, March V, Dong S, et al. Evaluation of a performance model of lustre file system[C]// 2010 Fifth Annual ChinaGrid Conference. 2010: 191-196.文件系统作为存储系统，存储容量为 100TB。存储架构为一个存储节点与两个扩展盘柜，扩展盘柜和中央存储节点直接使用 SAS 接口连接，使用回环模式。Lustre 文件系统是一个开源的，基于对象存储技术的集群并行文件系统，可为 Taurus 提供可靠、安全、易用且可扩展的存储环境 Kosta L, Hunter H, George G, et al. Measuring I/O Performance of Lustre and the Temporary File System for Tradespace Applications on HPC Systems[C]//Proceedings of the SouthEast Conference. 2017: 187-190.。Lustre 文件系统的架构图如图 2。

图 2 Lustre 文件系统架构图

Fig.2 File system architecture of Lustre

Taurus 存储使用集中管理方式，所有计算节点加载同一套存储系统。为了解决故障及 I/O 瓶颈问题 Yildiz O, Dorier M, Ibrahim S, et al. On the root causes of cross-application I/O interference in HPC storage systems[C]// 2016 IEEE International Parallel and Distributed Processing Symposium. 2016: 750-759.，存储控制器上两个独立的控制卡分别通过光纤和两个互为冗余的 I/O 节点连接，当其中一个 I/O 节点发生故障，另一个 I/O 节点立刻接管，保证了存储的稳定性。其次 I/O 节点和服务器之间采用 56Gb InfinibandPfister G F. An introduction to the infiniband architecture[M]// High Performance Mass Storage

and Parallel I/O: Technologies and Applications . 2001: 617-632.链接，支持多并发链接转换技术，具备高速数据传输能力。

通过文件系统基准测试工具 IOZone 对 Taurus 的文件系统进行了写入、读取、重读、重写、反向读、跨越式读、从文件中读、往文件中写、随机读取、预读取、内存映射文件 I/O、异步 I/O 读取、异步 I/O 写入等测试张春明，芮建武，何婷婷. 一种 Hadoop 小文件存储和读取的方法[J]. 计算机应用与软件，2012, 29(11): 95-100.。测试中指定生成的测试文件应小于实际内存容量，否则将影响测试结果。Taurus 高性能计算系统多节点最高读取速度达到 6GB/s 左右；写入速度如图 3，当文件大小为 256MB 且分块大小为 16MB 时达到最快的写入速度 650MB/s。

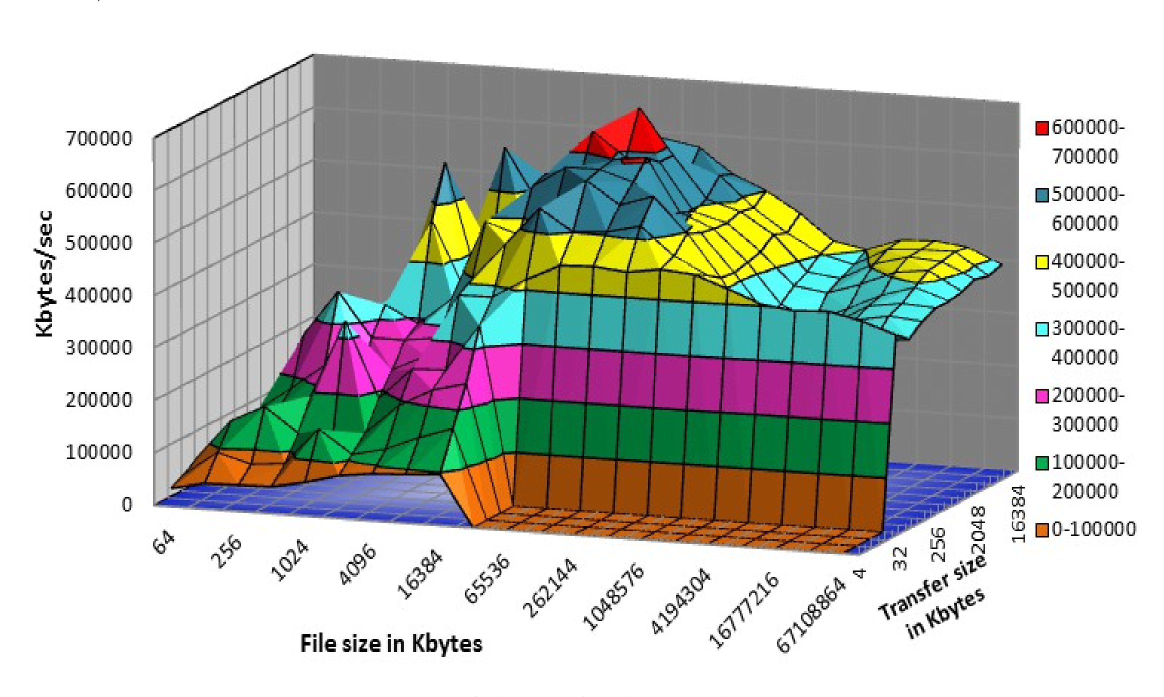


图 3 Taurus 高性能计算系统写入性能测试

Fig. 3 Write performance test of Taurus

1.4 软件环境

Taurus 高性能计算系统采用 PBS 集群调度器进行硬件资源的管理和分配，用户可以通过 PBS 提交作业脚本、查询作业状态信息、作业队列管理等操作。Taurus 采用 module 管理软件版本及库函数版本等，方便用户配置环境变量。用户可以直接使用 module 切换 python 版本、MPI 并行环境及编程模型等。 GPU 编程方面，每个计算节点支持 CUDA 运算架构，可以直接通过 PBS 提交数据密集型作业给 GPU 进行运算。

2、高性能计算系统应用

Taurus 高性能计算系统建成后，30 余名科研人员使用 Taurus 开展科研相关计算工作。

2.1 蒙特卡罗模拟分子云中的化学演化

天体化学是一门模拟各种各样分子在分子云中合成的学科，蒙特卡罗模拟是天体化学的一种常见模拟方法Lamberts T, Cuppen H M, Ioppolo S, et al. Water formation at low temperatures by surface O 2 hydrogenation III: Monte Carlo simulation[J]. Physical Chemistry Chemical Physics, 2013, 15(21): 8287-8302.。蒙特卡罗模拟方法是一种随机过程，用来模拟一个反应网络中某一个化学反应的发生。新疆天文台的天体化学课题组使用 Taurus 模拟在一个有尺度分布的系统中各种分子的化学演化，取分子云中的一个很小的体积作为一个系统，里面包含一个尘埃以及它周围的气体。这个系统中包含多种化学反应，可以通过计算系统对这些化学反应进行模拟，同时考虑多个尘埃时需并行共同演化，当达到一定条件后再对各个子系统进行混合处理，保证整个大系统处于一种均匀状态。

蒙特卡罗方法的缺点之一是耗时太长，在普通单机计算机上进行模拟二十万年演化时间尺度就需要数十天时间。而通过使用 Taurus 高性能计算系统，目前模拟二十万年演化时间尺度的时间约为 7 天。新疆天文台天体化学课题组使用 Taurus 高性能计算系统对不同分子演化使用蒙特卡罗方法模拟结果如图 4。

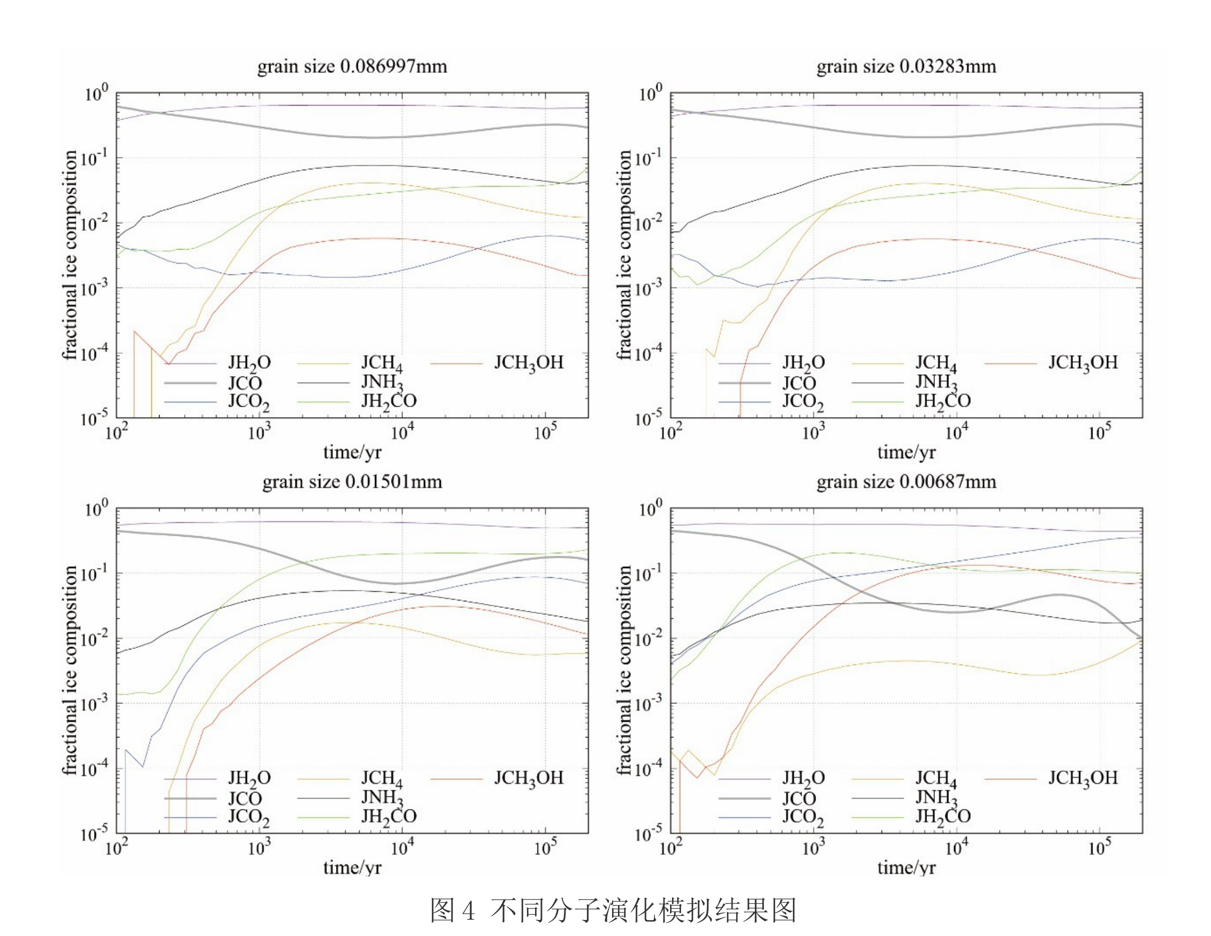


图 4 不同分子演化模拟结果图

Fig.4 Different molecular evolution simulation results

2.2 多相滤波及消干扰 GPU 加速

射频干扰(Radio Frequency Interference,RFI)的识别及消除，如何快速准确地找出射频干扰，防止把真实信号误判为干扰是一项急需解决的技术难点。由于大口径望远镜数据的计算量非常大、射频干扰环境复杂，对射频干扰实时处理技术提出很大的挑战。新疆天文台研究人员正在实验使用 CUDA 加速消除射频干扰，目前已初步实现基于 Taurus GPU 的自适应射频干扰处理方法，并得到良好的效果，射频干扰处理结果如图 5。使用 Taurus 高性能计算系统能够有效减少干扰处理消耗时间，为相干消色散实现提供了硬件平台。

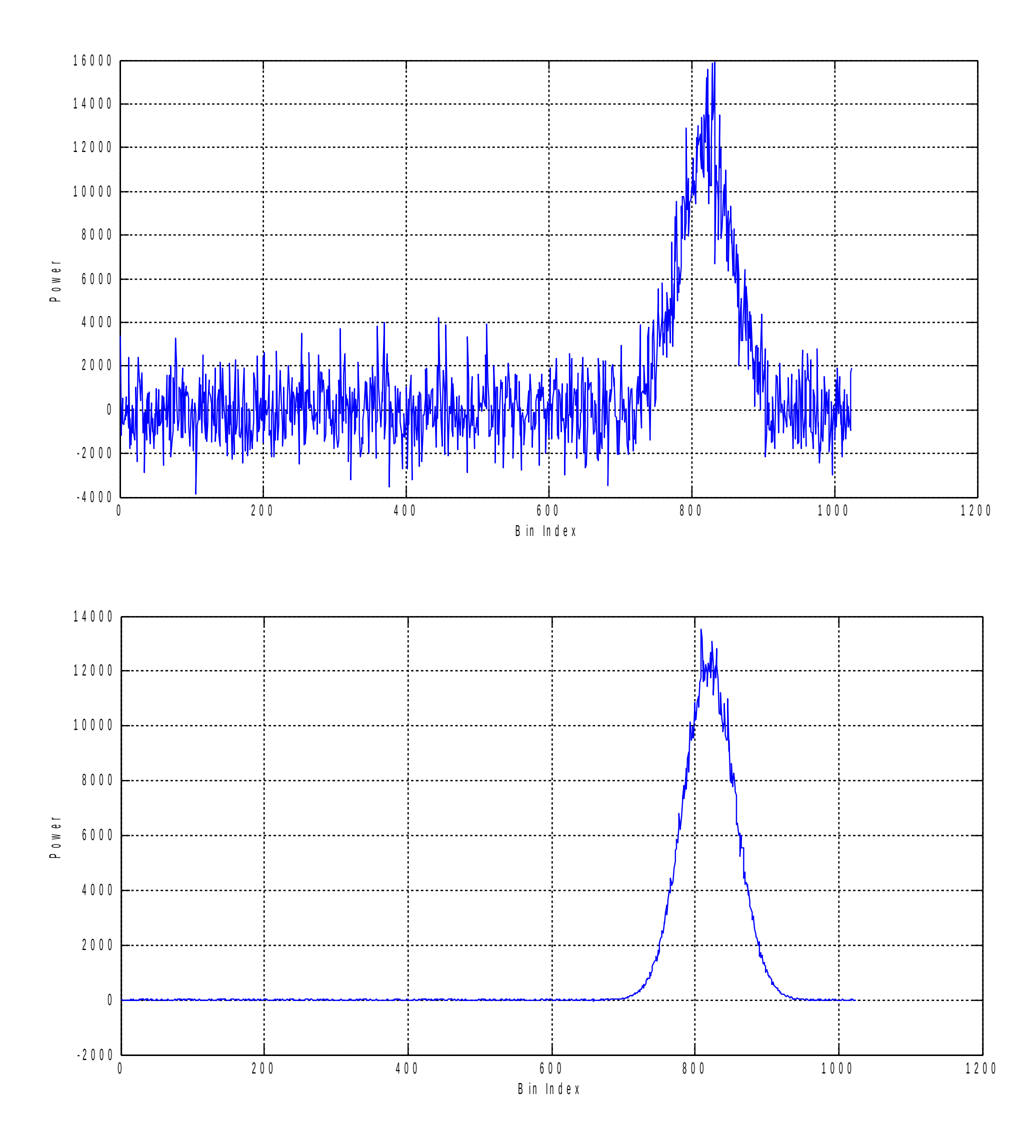


图 5 利用 Taurus GPU RFI 处理前、后脉冲星信号轮廓对比

Fig.5 The contrast of pulse profile by using Taurus GPU RFI

2.3 Taurus 高性能计算系统使用申请

Taurus 高性能计算系统采用开放式管理，接受相关领域研究人员申请，可将具体需求发送到 zhanghailong@xao.ac.cn。Taurus 平台详细使用说明及示例程序参见网站：http://taurus.xao.ac.cn。

3、结论

根据新疆天文台科研计算需求建设了 16 节点的高性能计算系统。经过测试，所建设的超算系统 CPU 性能为理论值的 93%，GPU 性能为理论值的 80%。Lustre 存储系统多节点在分块大小为 16MB、分块文件大于 256M 时取得较理想的写入速

度，如果文件过小影响整个存储系统性能。目前 30 余位用户在使用 Taurus 超算系统进行科学计算工作，在蒙特卡罗模拟分子云中的化学演化、多相滤波器组算法 GPU 加速等相关领域得到了较好的计算结果。

参考文献：

[1] 杨哲睿, 高娜, 刘梁. 大规模天文数据分析及多维信息可视化平台的建设和管理[J]. 科研信息化技术与应用, 2015, 6(5): 73-83.
Yang Zherui, Gao Na, Liu Liang. Construction and management of large scale astronomical data analysis and multi-dimensional information visualization platform[J]. E-science Technology & Application, 2015, 6(5): 73-83.

[2] Fan Z, Qiu F, Kaufman A, et al. GPU cluster for high performance computing[C]// Proceedings of the ACM/IEEE SC2004 Conference. 2004: 47-47.

[3] Kirk D. NVIDIA CUDA software and GPU parallel computing architecture[C]// Proceedings of the 6th international symposium on Memory management. 2007: 103-104.

[4] Chervenak A, Foster I, Kesselman C, et al. The data grid: towards an architecture for the distributed management and analysis of large scientific datasets[J]. Journal of Network and Computer Applications, 2000, 23(3): 187-200.

[5] Dongarra J J, Luszczek P, Petitet A. The LINPACK benchmark: past, present and future[J]. Concurrency and Computation: Practice and Experience, 2003, 15(9): 803-820.

[6] Barrett R F, Chan T H F, D'Azevedo E F, et al. Complex version of high performance computing LINPACK benchmark (HPL)[J]. Concurrency and Computation: Practice and Experience, 2010, 22(5): 573-587.

[7] Zhao T, March V, Dong S, et al. Evaluation of a performance model of lustre file system[C]// 2010 Fifth Annual ChinaGrid Conference. 2010: 191-196.

[8] Kosta L, Hunter H, George G, et al. Measuring I/O Performance of Lustre and the Temporary File System for Tradespace Applications on HPC Systems[C]//Proceedings of the SouthEast Conference. 2017: 187-190.

[9] Yildiz O, Dorier M, Ibrahim S, et al. On the root causes of cross-application I/O interference in HPC storage systems[C]// 2016 IEEE International Parallel and Distributed Processing Symposium. 2016: 750-759.

[10] Pfister G F. An introduction to the infiniband architecture[M]// High Performance Mass Storage and Parallel I/O: Technologies and Applications . 2001: 617-632.

[11] 张春明, 芮建武, 何婷婷. 一种 Hadoop 小文件存储和读取的方法[J]. 计算机应用与软件, 2012, 29(11): 95-100.
Zhang Chunming, Rui Jianwu, He Tingting. An approach for storing and accessing small files on Hadoop[J]. Computer Applications and Software, 2012, 29(11): 95-100.

[12] Lamberts T, Cuppen H M, Ioppolo S, et al. Water formation at low temperatures by surface O 2 hydrogenation III: Monte Carlo simulation[J]. Physical Chemistry Chemical Physics, 2013, 15(21): 8287-8302.

Taurus High Performance Computing System of Xinjiang Astronomical Observatory

Zhang Hailong[1,2], Ye Xinchen[1], Nie Jun[1,2], Chen Longfei[1], Tohtonur[1],Wang Jie[1], Cui Chenzhou[3], Li Changhua[3], Zhu Yan[1,4] Zhang Meng[1,4]

(1. Xinjiang Astronomical Observatory，Chinese Academy of Sciences，Urumqi 830011，China; 2. Key Laboratory of Radio Astronomy，Chinese Academy of Sciences，Nanjing 210008，China; 3. National Astronomical Observatories，Chinese Academy of Sciences，Beijing 100012，China; 4.University of Chinese Academy of Sciences,Beijing 100049,China)

Abstract: Taurus high performance computing system of Xinjiang Astronomical Observatory has 1 login node, 16 compute nodes, 2 I/O nodes and 100TB high-speed storage. In theory, the double precision floating-point computation capacity of CPUs is 6.7584Tflops. The actual peak turns out to be 6.289Tflops tested by Linpack, the available computation capability is 93.06% of the theoretical value. The computation capacity of the GPUs is 18.72Tflops in theory, while its practical peak is 14.882Tflops, the available computation capability is 79.5% of the theoretical value. The calculation nodes and the storage nodes are connected by 56Gb Infiniband network. Using IOZone for testing the storage performance, single-node writing reaches 460MB/s and multi-node writing can be 800MB/s. The Taurus HPC system has been applied in various fields such as GPU algorithm acceleration and Monte Carlo simulation.

Key words: HPC；Lustre；IOZone；